Identifying Emerging Safety Threats through Topic Modeling in the Aviation Safety Reporting System: A COVID-19 Study

Carlos Paradis
and Rick Kazman
University of Hawaii at Manoa
Honolulu, Hawaii, 96822, USA
Email: cvas@hawaii.edu
and kazman@hawaii.edu

Misty D. Davies
and Becky L. Hooey
NASA Ames Research Center
Moffett Field, CA, 94035, USA
Email: Misty.D.Davies@nasa.gov
and becky.l.hooey@nasa.gov

***Abstract*—In this work, we propose a method to identify emerging safety threats in the NASA's aviation safety reporting system using topic modeling. The proposed method serves both to connect different topics over time, but also to facilitate quick exploration and navigation. We use COVID-19 as a use case to identify if emerging safety threats can be identified *over time*. We found not only the topics' top terms indicated the prevalence of COVID-19 reports, but also they accounted for 71% of manually curated reports, offering an automated means to identify emerging threats.**

I. Introduction

The NASA Aviation Safety Reporting System (ASRS) is a voluntary confidential aviation safety reporting system. The ASRS receives reports from pilots, air traffic controllers, flight attendants and other involved in aviation operations. The reports are de-identified and coded by ASRS expert safety analysts and a short descriptive synopsis is written to describe the safety issue. The de-identified reports are then disseminated to the aviation community in a number of ways including entry into an online database, Safety Alert Bulletins and For Your Information Notices, and the CALLBACK newsletter. A quarterly report of ASRS noted that although its designers were acutely aware of the potential value of data derived from individual occurrences in highlighting deficiencies and discrepancies in the national aviation system, they also believe that other, perhaps more valuable, insights into system problems can be gained only by study of a large body of occurrence data [1].

In this work, we propose one approach for evaluation of a large body of occurrence data, by automatically identifying the evolution of ASRS report topics. We evaluate our proposed method using the recent surge of COVID-19 ASRS reports since March/April of 2020. Little metadata is available on these reports, and therefore we leverage only the ASRS report narratives to evaluate our approach. Specifically, we pose the following overarching research question:

> Can we detect the emerging COVID-19 topic timelines?

In the context of this work, *topics* are expected to reflect the subject of similar narratives, but also are mathematically defined in the literature on topic modeling [2]. In this research, we used the WarpLDA implementation for topic modeling [4]. This allows for the construction of timelines. We can not only observe if a topic is 'new/not new' but also what topics precede and succeed it over months, i.e. a topic timeline [6]. For the proposed method to be useful, we would like a topic model system that, taking as input narratives, identifies useful topic timelines to facilitate the exploration of ASRS narratives.

We chose COVID-19 ASRS reports for a few reasons. First, this was a very relevant theme as of the time of writing. Second, the vocabulary surrounding COVID-19 is not as specialized as other topics within ASRS, making a good candidate for evaluation in the scope of this work, as readers can easily relate to the identified themes. Lastly, ASRS has created a large report set for this theme, which also reflects its relevance to the organization and its audience, containing 1213 COVID-19 related reports to date, all of which are contained within the year of 2020, making a perfect candidate for evaluating our method. To empirically evaluate our overarching research question we define the following specific research questions:

> RQ1: Are there timelines in which the topic's top terms clearly suggest the prevalence of COVID-19 reports?

In topic modeling, topics are described as a sequence of words, known as 'top terms'. Topics also identify the original group of ASRS reports associated with it. Ideally, stakeholders should be able to navigate the groups of reports by understanding a topic's top terms. Here we are asking if the top terms from COVID-19 reports also display top terms that contain COVID-19 related words. If this is not the case

then users may not be able to find COVID-19 reports using topics.

RQ2: If there exist timelines in which the top terms suggest a prevalence of COVID-19 reports, what proportion of the COVID-19 Report Set do they account for?

Our goal here is to understand how many COVID-19 reports from the *COVID-19 Report Set* a user would have detected via the top terms in a given timeline, i.e. the precision and recall of the *report set* in a given timeline. We emphasize the report set here, because the COVID-19 Report Set do not contain every report of of 2020. We find nonetheless the analysis pertinent, as our goal is to automate the identification of emerging safety threats, and thus we would like to know the following: if the Report Set did not exist, how many of the reports in the identified timelines would have been identified.

RQ3: Are there timelines which consist exclusively of COVID-19 Reports?

Here we inquire into a slightly different question than RQ2: Rather than focusing on the proportion of reports from the COVID-19 Report Set, we focus on the proportion of said COVID-19 Reports per Timeline, i.e. how "saturated" is each identified timeline with COVID-19 reports from the Report Set. Ideally, we would like to know, for all COVID-19 reports in 2020, how many each timeline has. But this information is not available (as it would have required the exhaustive manual inspection of each incoming report). Nonetheless, if we can identify timelines that contain close to or 100% of the COVID 19 reports from the report set, then the effect is the same. Unique COVID-19 timelines are highly valuable from a user standpoint, as the user will have high confidence that, by inspecting such timelines, they will identify relevant COVID-19 reports.

II. Method

We begin describing our method by defining timelines, building on the work of Smith et al. [6]. Our implementation is available as part of the R package kaona[1]. In Figure 1, each rectangle represents a topic. This figure also illustrates that each month may have multiple topics. In this paper, our focus is in empirically deriving the *edges* between the rectangles shown—to create a timeline—and devising a method of evaluation to assess if the identified topic timeline is correct.

A. Constructing Monthly Topics

To obtain multiple topics *per month*, as shown in Figure 2, we downloaded all ASRS 2020 reports publicly available via the ASRS Database Online[2] in tabular format.

[1]http://github.com/sailuh/kaona

[2]https://asrs.arc.nasa.gov/search/database.html

Each ASRS report contains a timestamp of the format year-month (day or time information is not available). We used this timestamp to partition the entire collection of 2020 reports in months, and then applied WarpLDA to each month *separately*. Therefore, in the example shown in Figure 1, which displays six months, WarpLDA would have been applied six times, once for each month. The output of WarpLDA is the topics, i.e. the rectangles shown in the figure. In actuality, ASRS has available at the time of the writing eleven months worth of reports (January through November), and therefore WarpLDA was applied eleven times in this work.

An important consideration here is the number of topics, k, which in turn reflects the number of rectangles shown in Figure 1 per month, as the number of topics is not obtained automatically. Since the algorithm is executed independently per month, any number of topics ranging from 2 and up is a candidate number for k; we chose k = 10 for all months.

According to [3], utilizing metrics to optimize for perplexity groupings led to less meaningful sets of words describing topics presented to users. A method to address the conflict, however, is not proposed. This is consistent with the closest work related to ours in [5] which applies topic modeling to identify trends in a NASA dataset. The authors state that the process of selecting the topic modeling parameters *lack* definitive guidance. We therefore decided to emphasize performance instead: The number of topics is reasonable sized to not be too short or too computationally expensive to our following step of identifying candidate edges among the resulting 11 months * 10 = 110 topics.

We chose this number for performance purposes only, as the construction of the edges, as we will discuss next, requires the Cartesian product of every pair of consecutive month topics to assess their similarity, and to construct the entire year's timeline requires the construction of 110 topics.

Finally, we expect our proposed method, a variant of *Topic Flow* [6] , to compensate for any underestimate or overestimate of the "true" number of topics in a given month: If a given month has too many redundant topics, we expect them all to converge to a single topic in the following month. Conversely, if there are too few topics in a month, we expect them to split in different topics in the following month.

B. Connecting Topics over Time

As we noted earlier, each rectangle in Figure 1 is a topic, which in turn are described by set of words. Specifically, all topics (rectangles) are represented by topic-term matrices: Each rectangle is a row in a table, and each column is a word. The value of the cell is a probability assignment from a word to a topic. Choosing the highest probability words for each topic is the standard practice to choose the words to describe a topic.

Specifically, after we construct the monthly topics in Figure 2, all topic's (rectangles) topic-term matrices in month *i* are paired (via Cartesian product) with month *i+1* (e.g. Jan-Feb, Feb-Mar, etc.). For each pair we then compute the cosine

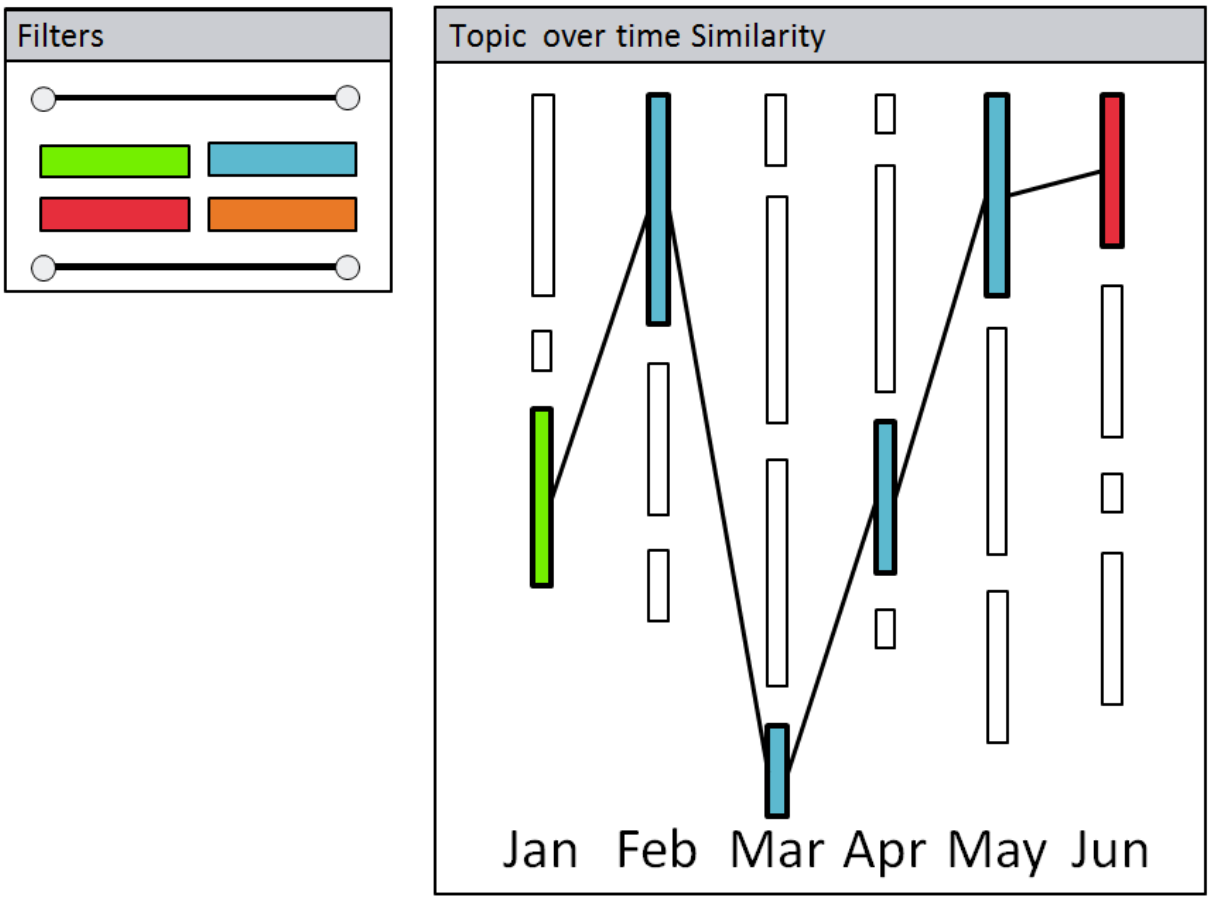

Fig. 1. Topic Flow Model, adapted from [6].

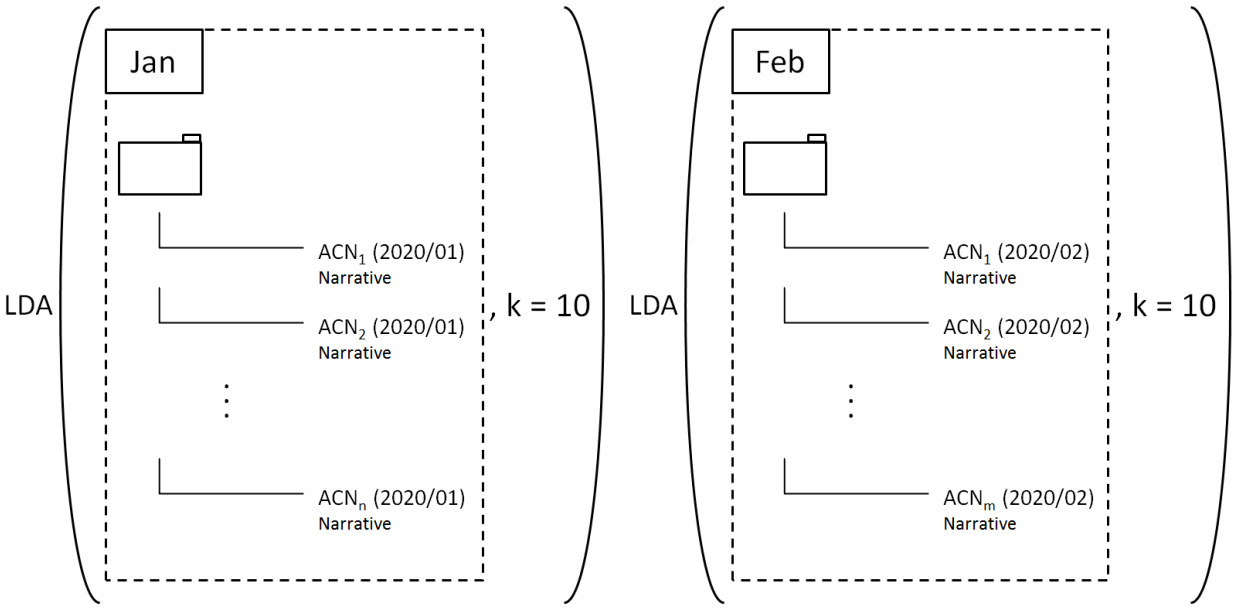

Fig. 2. ASRS Report Partition for Monthly Topics.

similarity of the topic-term matrices, as defined in 1. The Cartesian product is shown in Figure 3.

$$\cos(\theta) = \frac{z_i \cdot z_j}{\|z_i\|_i \|z_j\|_j} \tag{1}$$

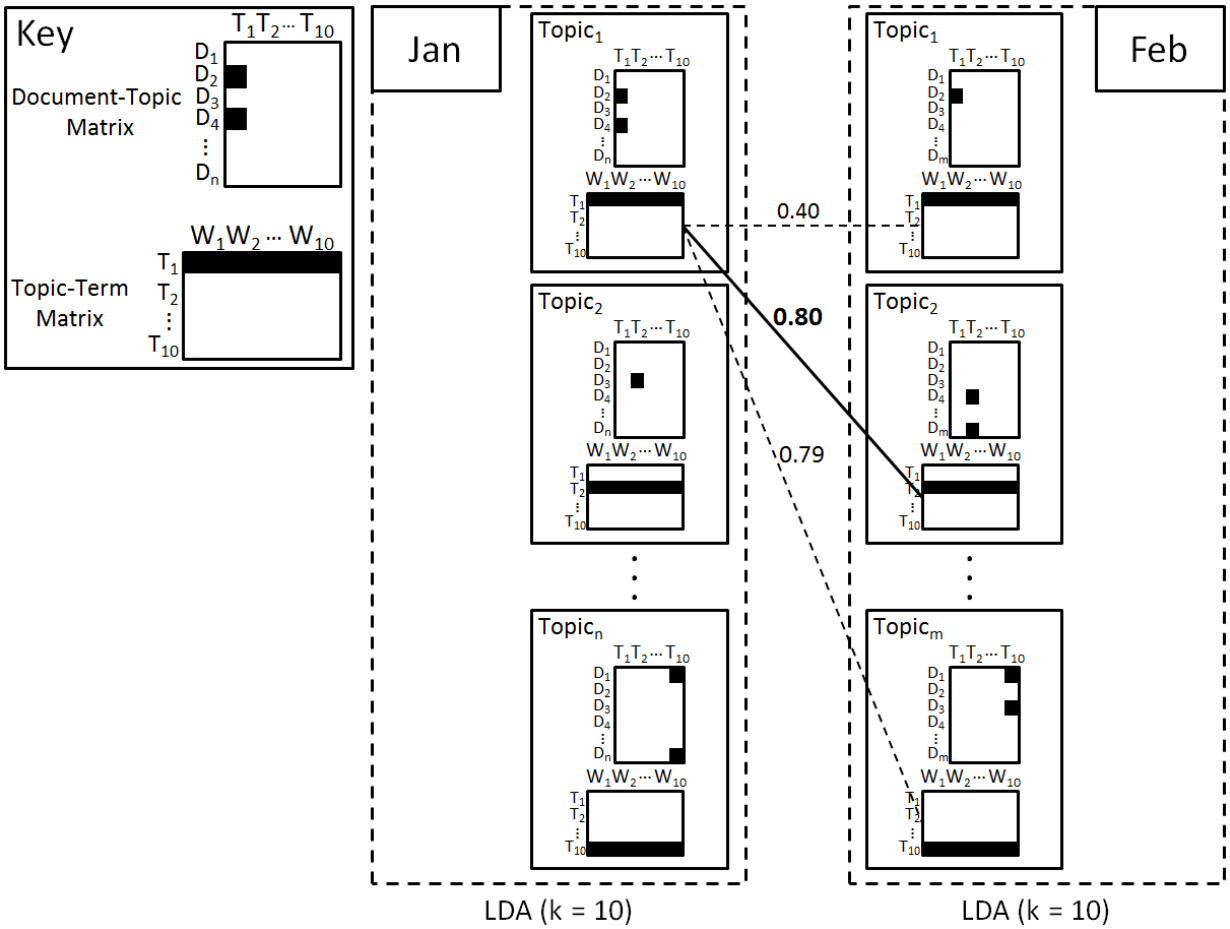

Fig. 3. Connecting topics over time using topic term matrices.

where in equation 1, z_i and z_j represent topic-term matrices. The intuition here is that, if the distribution of the set of words are similar between topics, then they are considered similar.

For example, for topic 1 in January, its similarity to all 10 topics in February is calculated. What remains then is to decide which candidate edges to keep, resulting in a diagram similar to the one shown in Figure 1.

In [6], the authors apply a global threshold to all topics, i.e. a user must use a slider (top left of the Figure 1) to adjust what threshold should be used between all topics to preserve the lines. We see some limitations with this approach: The decision of threshold is entirely up to the user by trial and error, and it is not necessarily clear how the user would have achieved the correct parameter. Moreover, we find it unlikely that a single global threshold can effectively reflect all pairs of topics, and believe a different approach to establish the line filtering criteria could more accurately reflect the trends. Specifically, we chose to use maximum likelihood instead, and include alongside each edge the associated similarity weight (which ranges from 0 to 100%).

Consider again the previous example where January's topic 1 has 10 candidate edges to the 10 topics of February. Suppose also among the 10 February topics, Topic 7 has the highest similarity to January's topic 1. Then, via maximum likelihood, we claim topic 1 is connected to topic 7, and the remaining candidate topic edges are deleted. A major threat to validity in this case, is that topics may end up connected even though they have very low similarity. However, since we know the similarity weights, while inspecting the terms of January's topic 1, the terms of February topic 7, and the similarity weights of other topic pairs in the two months, we may conclude that the topics are in reality disjoint. If this is the case, then all the remaining 9 candidate edges would also be disjoint, since the chosen edge had the highest similarity weight. Therefore, we find this an easier approach to derive the edges from the candidate edges in Topic Flow.

C. Paths and Timelines

To facilitate discussion in the following results section, we define the set of *single* topics (rectangles) shown in Figure 1 connected across all months displayed as a *path*, in the sense of graph theory. A timeline is defined as a set of paths that contain at least one topic in common.

D. Types of Topics within Timelines

In original definition of Topic Flow [6], the lines which connect the rectangles characterize the rectangles in one of

four types: emerging, continuing, ending and standalone. In Figure 1, the green rectangle represents *emerging*, the blue rectangle *continuing*, and the red rectangle *ending* topics. A standalone orange rectangle would resemble the colorless rectangles shown in the Figure, although we left them colorless to emphasize the other colored three, which form a *a timeline*. Specifically, an emerging topic has no earlier month's topics connecting to it. Conversely, an ending topic has no later month's topics connecting to it. Continuing topics are those that have both prior and following month edges, and standalone topics are those that have neither.

In the maximum likelihood criteria we adopt, Topic Flow will generate all but standalone topics. However if, as we exemplified earlier, users deem for any pair of topics the similarity weight to be low, and/or the set of terms of the chosen month pair disjoint, then in this case the topic may become standalone.

E. Deterministic Mapping Criteria

To answer RQs 2 and 3, we must determine whether documents belong or not to a given to a given topic, and therefore to a given path and timeline. That is, we must provide a deterministic mapping to evaluate if the topic assignment is accurate using our evaluation dataset. We used a maximum likelihood assumption, i.e., we assume the highest likelihood reflects the more appropriate, and importantly, single mapping.

III. Results

The results of performing the Topic Flow analysis over the entire 2020 ASRS Database are split into Figures 4 and 5 due to space constraints. We now answer our three research questions.

> RQ1: Are there timelines in which the topic's top terms clearly suggest the prevalence of COVID-19 reports?

We manually identified, using *only the terms of each topic*, a total of 19 COVID-19 topics out of the 110 topics, distributed across 6 paths (Path IDs 12, 4, 2, 6 1 and 32) and 4 timelines (1, 5, 6, 32). The paths are shown in Figures 4 and 5. The list of terms associated with each of the 19 identified topics is shown in Tables I and II. The identified COVID-19 topics are also highlighted in a red rectangle in Figures 4 and 5.

Therefore, the answer to RQ1 is *Yes*, it is possible to identify timelines in which the terms clearly suggest the prevalence of COVID-19 reports. We note the first COVID-19 related topic through the set of terms dates from March, which is the first occurrence of COVID-19 reports in the ASRS COVID-19 Report Set (Timeline 1, Path ID 12 in Figures 4,5. We also observed from the Figure that at least one topic in each month of the timeline from March onward contains terms associated with COVID-19, but in different timelines. We would expect therefore that they convey different meanings, or otherwise the topics should have been part of the same paths.

In *Path ID 12*, we see terms suggestive of crew preparedness through company training to assist passengers due to the virus. *Path ID 4* emphasizes the use of face masks and row seats for passenger safety. The focus of *Path ID 6* is on the effect of COVID-19 on work days of the crew, and airport-related flying safety issues. Common to *Path IDs 4, 6, and 2* are the topics in October and November, which are associated with company training for COVID-19 situations, which is a reasonable common ground for all three path ids to converge.

The *Path IDs 1 and 32* convey the use of masks and cleaning and shift procedures for COVID safety for controllers. While both *Path ID 1* and *Path ID 4* are associated with mask wearing, *Path ID 4* seems related to boarding and social distancing due to policy (terms: gate, policy, boarding, policy, social, distancing) whereas in *Path ID 1* mask wearing concerns appear more associated to the positioning of passengers on the plane (terms: rows, seat, passengers, service, takeoff).

While the full interpretation of the terms is open for debate, we believe it is reasonable from the presented list of terms that they are suggestive of different COVID-19 themes in different path ids, and suggestive of similar themes, *within* path ids, as we would intuitively hope they would be.

We emphasize that the COVID-19 related topics appear consecutively, instead of scattered across Figures 4,5, consistent with the intuition of emerging timelines, and with the effect of cosine similarity on topic-term matrices of consecutive months.

> RQ2: If there exist timelines which the top terms suggest prevalence of COVID-19 reports, what proportion of COVID-19 reports of the Report Set do they account for?

In Figure 4,5, if a user were to have inspected the *entire timeline* where at least one COVID-19 topic occurs (based only on the use of terms we chose), then they would have encountered approximately 14.8% + 29.0% + 26.3% + 0.9% = 71% of the COVID-19 reports from the COVID-19 report set. Notice that, in doing so, they would have disregarded the timelines with the fewest of the COVID-19 report set's reports (i.e. timelines 2, 3, 4, 7, 8 and 9, with the exception of timeline 10).

However, this is not the only approach a user may choose to navigate the reports. Instead, the search can be constrained to only the paths where a COVID-19 topic occur, or to the topics themselves in the interest of time. The ability to choose a trade-off between precision and recall here, accounting for human effort, rather than only keyword searching, we believe, highlights the benefit of our proposed method.

> RQ3: Are there timelines which consist exclusively of COVID-19 Reports?

In Figures 4 and 5, we can see a wide variety of saturation of COVID-19 reports from the report set, ranging from timelines 6 (92.5%) to timeline 3 (12.6%). However, there are no

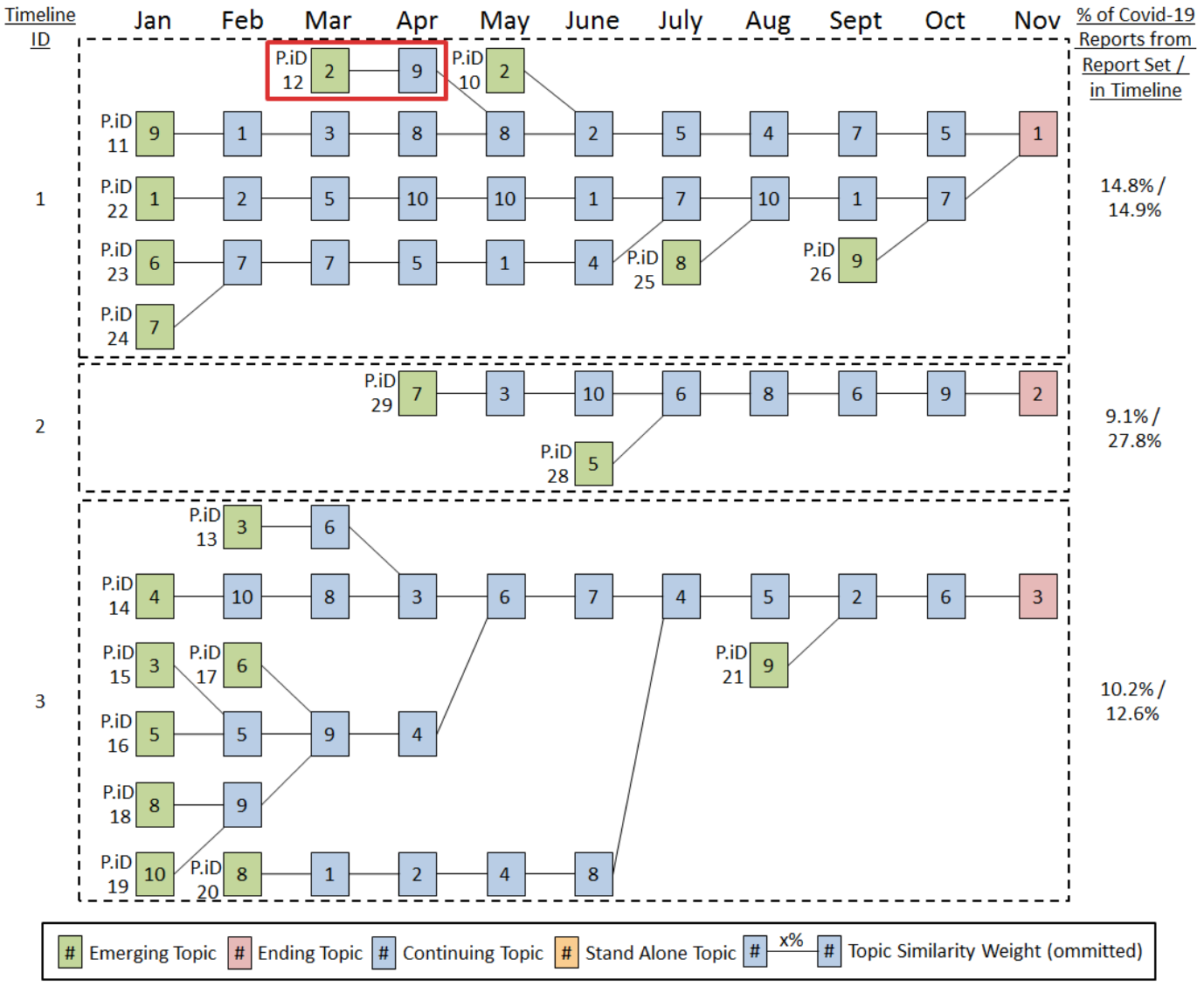

Fig. 4. ASRS 2020 Topic Flow. Path IDs are abbreviated as P.iD.

exclusive COVID-19 timelines, and hence the answer to RQ3 is no.

A lower saturation of COVID-19 reports here suggests non-related COVID-19 themes which may overlap in problem, equipment or solution to COVID-19 reports. For instance, in Topic 5 we can see path ids 4, 2, and 6 have COVID-19 related topics (per their display of terms). However, July's topic 3, and August's topic 3 are not. The overlap, however, occurs on issues during clearance and takeoff (terms: tower, clearance, departure), which is also the overarching topic for path id 2 (terms: tower, approach, VFR, controller taxiway).

IV. Conclusion and Future Work

In this work, we combined methods from prior research to define a variant of Topic Flow. We provided both qualitative and qualitative evaluation of the viability of Topic Flow capturing emerging safety threats using a timely emerging theme. Going beyond prior research in this area, we considered the results of Topic Flow in the context of a real world system and the merit of its usage in operation.

Our construction of the Topic Flow method allows for the identification of emerging safety threats, as it does not attempt to fit the entire year of topics a priori. Rather, a timeline can be defined as soon as any time granularity (e.g. month, week) is available, and also allows for the existence of topics that did not exist up to a given point to emerge.

Acknowledgments

The material is based upon work supported by NASA under award No 80NSSC19M0202. This research was partially conducted at NASA Ames Research Center. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not

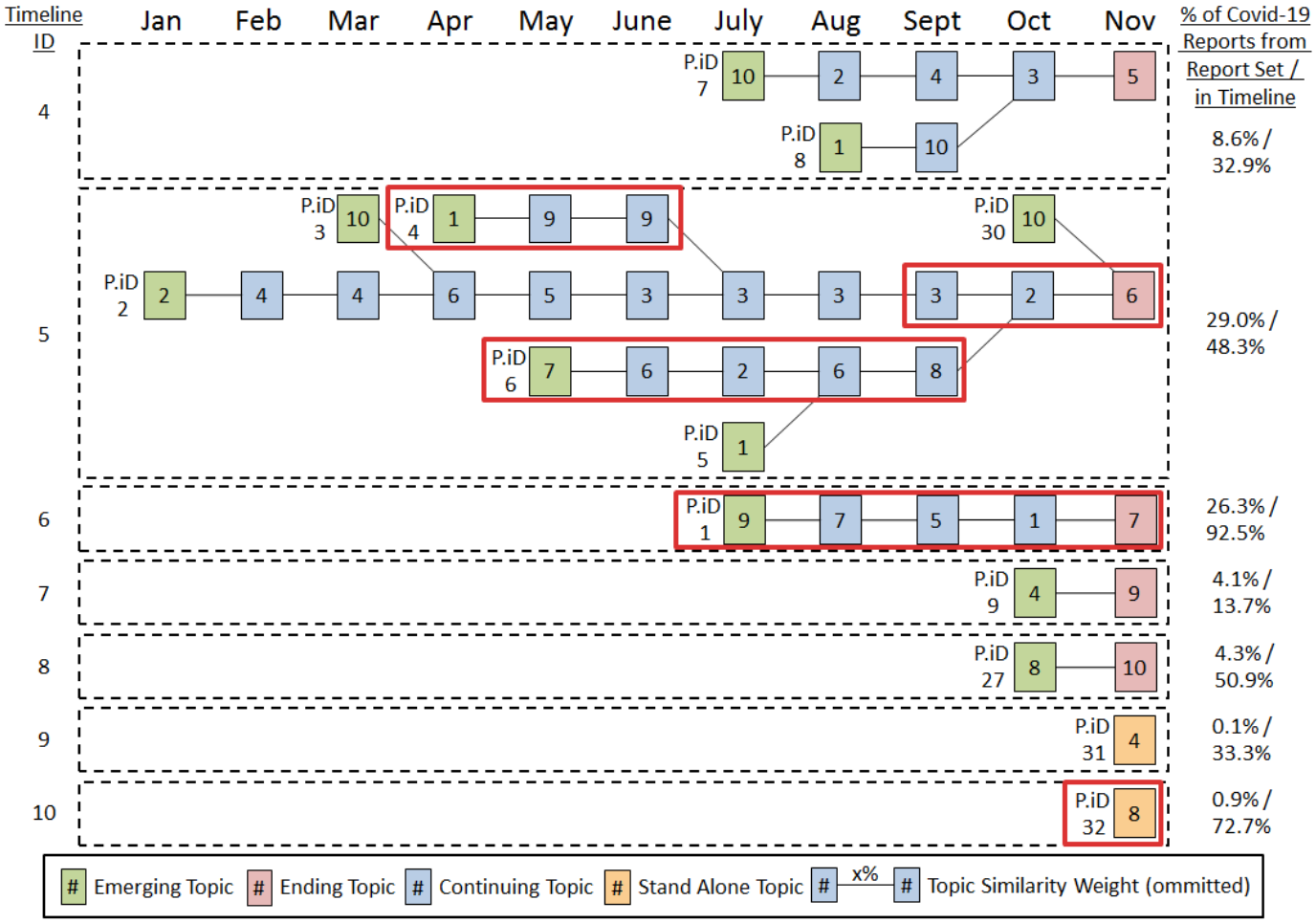

Fig. 5. ASRS 2020 Topic Flow (Continued). Path IDs are abbreviated as P.iD.

constitute or imply its endorsement by the United States Government.

References

[1] CE Billings, JK Lauber, H Funkhouser, EG Lyman, and EM Huff. Nasa aviation safety reporting system. 1976.

[2] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 2003.

[3] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc., 2009.

[4] Jianfei Chen, Kaiwei Li, Jun Zhu, and Wenguang Chen. Warplda: a simple and efficient o(1) algorithm for latent dirichlet allocation. *CoRR*, abs/1510.08628, 2015.

[5] Lucas Layman, Allen P. Nikora, Joshua Meek, and Tim Menzies. Topic modeling of nasa space system problem reports: Research in practice. In *Proceedings of the 13th International Conference on Mining Software Repositories*, MSR '16, page 303–314, New York, NY, USA, 2016. Association for Computing Machinery.

[6] A. Smith, S. Malik, and B. Shneiderman. *Visual Analysis of Topical Evolution in Unstructured Text: Design and Evaluation of TopicFlow*, pages 159–175. Springer, 2015.

Timeline ID	Path ID	Topic ID	Month	Top 10 Terms
5	**2**	**2**	**Jan**	**traffic tower approach airport clearance pilot final departure turn frequency**
5	**2**	**4**	**Feb**	**traffic tower final cleared told left pattern pilot downwind made**
5	**2**	**4**	**Mar**	**tower turn airport miles fuel left landing approach flying frequency**
5	**2**	**6**	**Apr**	**tower class clearance traffic airspace vfr departure route called asked**
5	**2**	**5**	**May**	**tower taxiway taxi controller control cleared contact airspace plan short**
5	**2**	**3**	**Jun**	**tower controller clearance departure asked told ground called cleared frequency**
5	**2**	**3**	**Jul**	**tower taxi clearance ground frequency ramp taxiway turn controller cleared**
5	**2**	**3**	**Aug**	**departure clearance crew due pilots noticed area correct change takeoff**
5	2	3	Sep	due close pilots flying change company covid times missed atc
5	2	2	Oct	day due company covid hours captain event trip message flying
5	2	6	Nov	company training crew covid month due informed situation return pilots

TABLE I

SET OF TOP 10 RELATED WORDS FOR THE TOPICS OF PATH ID 2, WHICH CONTAIN COVID-19 RELATED TERMS. THESE TOPICS CAN BE SEEN IN THE FIGURE. THIS DISPLAYS ALL THE TOPICS IN PATH ID 2, BOTH COVID-19 RELATED AND NON COVID-19 RELATED WORDS. ROWS IN BOLD INDICATE ADJACENT TOPICS IN WHICH TERMS WERE DEEMED NOT COVID-19 RELATED.

Timeline ID	Path ID	Topic ID	Month	Top 10 Terms
1	12	2	Mar	company safety supervisor told covid passengers training virus crew hand
1	12	9	Apr	crew covid company information due informed action day safety fact
1	**12**	**8**	**May**	**maintenance crew fuel made cockpit ramp position make situation dangerous**
5	4	1	Apr	mask passengers passenger safety masks covid fa rows people seat
5	4	9	May	passenger mask masks safety face crews service told wearing seat
5	4	9	Jun	taxi taxiway short line passenger hold ramp takeoff make mask
5	**4**	**3**	**Jul**	**tower taxi clearance ground frequency ramp taxiway turn controller cleared**
5	6	7	May	flying due part noticed day failed covid panel looked flown
5	6	6	Jun	day crew work days hours part event covid working company
5	6	2	Jul	airport working covid day area flying issue safety due information
5	6	6	Aug	day company crew told covid part number ramp work call
5	6	8	Sep	safety air day work made hours operations covid service make
6	1	9	Jul	mask passenger passengers asked fa seat told put gate policy
6	1	7	Aug	mask passenger passengers fa seat asked policy told face captain
6	1	5	Sep	mask passenger passengers asked fa told put policy wear attendant
6	1	1	Oct	mask passenger passengers attendant wear seat fa asked wearing boarding
6	1	7	Nov	passenger passengers policy mask fa seat attendant service distancing social
10	32	8	Nov	tower work flights shift cleaning windshear told controllers safety covid

TABLE II

SET OF TOP 10 RELATED WORDS FOR THE TOPICS OF PATH IDS 1, 4, 6, 12, AND 32, WHICH CONTAIN COVID-19 RELATED TERMS. THESE TOPICS CAN BE SEEN IN THE FIGURE SURROUNDED BY RED RECTANGLES. ROWS IN BOLD INDICATE ADJACENT TOPICS IN WHICH TERMS WERE DEEMED NOT COVID-19 RELATED.